EvalCrafter: Benchmarking and Evaluating Large Video Generation Models

Year: 2023 | Citations: 219 | Authors: Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang

---

Abstract

The vision and language generative models have been overgrown in recent years. For video generation, various open-sourced models and public-available services have been developed to generate high-quality videos. However, these methods often use a few metrics, e.g., FVD [56] or IS [45], to evaluate the performance. We argue that it is hard to judge the large conditional generative models from the simple metrics since these models are often trained on very large datasets with multi-aspect abilities. Thus, we propose a novel framework and pipeline for exhaustively evaluating the performance of the generated videos. Our approach involves generating a diverse and comprehensive list of 700 prompts for text-to-video generation, which is based on an analysis of real-world user data and generated with the assistance of a large language model. Then, we evaluate the state-of-the-art video generative models on our carefully designed benchmark, in terms of visual qualities, content qualities, motion qualities, and text-video alignment with 17 well-selected objective metrics. To obtain the finalleaderboard of the models, we further fit a series of coefficients to align the objective metrics to the users' opinions. Based on the proposed human alignment method, our final score shows a higher correlation than simply averaging the metrics, showing the effectiveness of the proposed evaluation method.